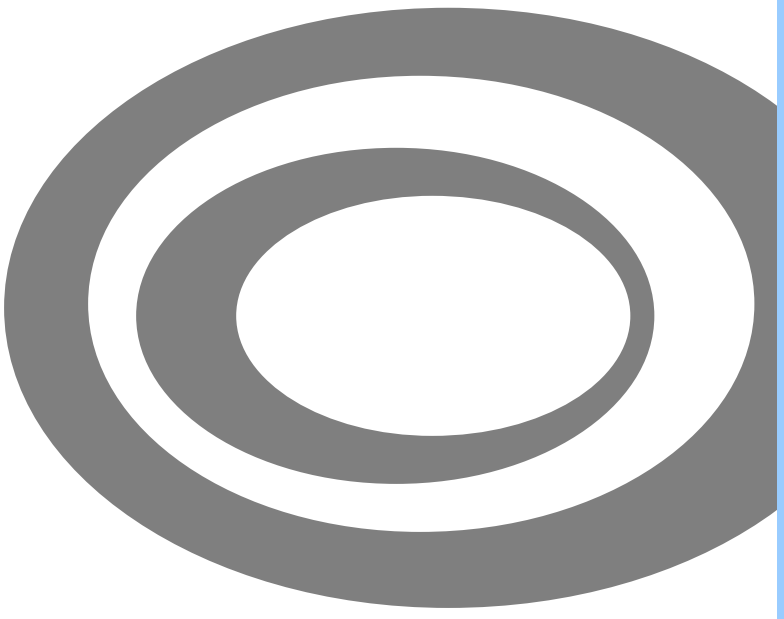
Programa de Promoción de la Reforma Educativa en América Latina y el Caribe

# Grupo de Trabajo sobre Estándares y Evaluación

# Validez de la interpretación de una prueba y su uso

**Samuel Messick**



Grupo de Análisis para el Desarrollo

**Validez de la interpretación y del uso de una prueba**

Samuel Messick[1]

*Educational Testing Service*

Tradujo del inglés: Hugo Mora Poltronieri, M.Sc. Universidad de Costa Rica

La validez es un juicio evaluativo integrado del grado hasta el cual, tanto la evidencia empírica como las justificaciones teóricas apoyan la adecuación y la pertinencia de las interpretaciones y las acciones basadas en los puntajes de pruebas u otros tipos de medición. Los principios de la validez se aplican no sólo a las inferencias interpretativas y de acción surgidas de los puntajes de pruebas tal y como son usualmente concebidos, sino también a inferencias obtenidas a partir de cualquier medio para observar o para documentar comportamientos o atributos consistentes.

Por consiguiente, en este documento el término *puntaje* se usará en su sentido más amplio, para hacer referencia a cualquier codificación o resumen de consistencias observadas, o de regularidades en el desempeño en una prueba, cuestionario, procedimiento de observación u otro medio de evaluación (tales como muestras de trabajo, portafolios[2] o simulaciones de problemas reales). Tal uso general incluye tanto resúmenes de tipo cualitativo como cuantitativo. Se aplica, por ejemplo, a protocolos, a interpretaciones clínicas, a juicios o valoraciones sobre comportamiento o desempeño, o a informes sobre puntajes verbales computarizados. En tal sentido amplio, los puntajes tampoco están limitados a atributos y consistencias del comportamiento humano, tales como la persistencia y la habilidad verbal. Los puntajes pueden referirse, asimismo, a consistencias funcionales y a atributos de grupos, situaciones o ambientes, así como a objetos o instituciones, como en el caso de mediciones sobre solidaridad grupal, estrés

[1] Este artículo aparece en M. C. Alkin, (Ed.), *Encyclopedia of Educational Research* (6th. ed.), New York: Macmillan, 1991. Se agradecen los valiosos comentarios de Walter Emmerich, Robert Linn y Lawrence Striker.

[2] Dentro del contexto de la medición, debe entenderse como una técnica empleada en evaluación del desempeño. Consiste en una colección de producciones o productos del estudiante en que quedan reflejados sus procesos y su rendimiento durante un período escolar dado. (N. del T.)

situacional, calidad de la producción artística, lo mismo que a indicadores sociales tales como la deserción escolar.

En sentido amplio, la validez es un resumen inductivo tanto de la evidencia existente, como de las consecuencias reales y potenciales de la interpretación y del uso de los puntajes. Visto así, lo que debe validarse no es la prueba o el instrumento como tal, sino las inferencias que pueden derivarse de los puntajes o de otros indicadores (Cronbach, 1971), es decir, las inferencias sobre lo que significan los puntajes o su interpretación, así como las implicaciones para la acción que dicha interpretación acarrea.

Es importante destacar que la validez es cosa de grado, no de "todo o nada". Además, con el tiempo, la evidencia de validez existente aumenta (o se desvirtúa) por nuevos hallazgos. Más aún: las proyecciones acerca de las posibles consecuencias sociales de las pruebas cambian a causa de evidencias sobre las consecuencias reales y por las condiciones sociales cambiantes. En principio, entonces, la validez es una propiedad en evolución; la validación, un proceso continuo, excepto, por supuesto, en el caso de las pruebas que obviamente son inadecuadas o inapropiadas para la interpretación o uso que se les quiera dar. En la práctica, dado que la evidencia de validez es siempre incompleta, de lo que se trata es de sacar todo el provecho posible de la situación, a partir de la evidencia disponible, tanto para justificar el empleo de la prueba en un determinado momento, como para guiar la investigación existente necesaria para mejorar la comprensión del significado de los puntajes, y para determinar cómo éstos pueden interpretarse dentro del contexto en que se produjeron. Esta investigación validatoria para ampliar la evidencia inmediata sirve luego, ya sea para corroborar o para revisar los juicios de validez obtenidos previamente.

Validar una inferencia interpretativa significa tomar en cuenta hasta qué punto una serie de líneas de evidencia es congruente con la inferencia misma; y, a la vez, determinar que todas las otras evidencias alternativas cuentan con un sustento menor. Se denomina evidencia

*convergente* a los resultados de investigación que dan apoyo a la interpretación propuesta o a un determinado uso de la prueba. Así, la evidencia convergente para una prueba de aritmética (en que los problemas se enuncian explícitamente por escrito), interpretada como una medida de razonamiento cuantitativo, podría indicar que los puntajes se correlacionan ampliamente con el desempeño en problemas lógicos, que distinguen o "discriminan" entre estudiantes orientados hacia la especialización en matemáticas y los que lo hacen hacia el inglés como carrera, y que esos mismos puntajes también son buenos predictores de éxito en cursos de ciencias. Los resultados de investigación que permiten descartar inferencias alternativas y, con ello, dar mayor credibilidad a la interpretación propuesta, se conocen como evidencia *discriminante*. Por ejemplo, para contrarrestar la posibilidad de que una prueba de problemas verbales sea en realidad una prueba disimulada de comprensión de lectura, se podría demostrar que las correlaciones con puntajes de lectura no son excesivamente altas, que el peso otorgado al factor de comprensión verbal es insignificante y que el nivel de lectura requerido por los ítemes no es difícil para la población a que pertenece el grupo de interés. Tanto la evidencia convergente como la discriminante cumplen una función fundamental en la validación de pruebas (Campbell & Fiske, 1959).

Para validar una inferencia que lleve a la acción se necesita validar no sólo el significado del puntaje, sino también las implicaciones de valor y las consecuencias de la acción, en especial la relevancia y utilidad de las consideraciones sobre los puntajes en propósitos aplicados, y las consecuencias sociales esperadas e inesperadas de usar los puntajes para tomar decisiones. Por ejemplo, supongamos que los puntajes de la prueba de problemas verbales antes mencionada –basados en evidencia convergente y discriminante- pueden realmente ser interpretados a partir del *constructo* conocido como razonamiento cuantitativo. En la literatura sobre validez, el término "constructo" se usa generalmente para hacer referencia al significado de los puntajes, -por lo general, aunque no necesariamente, atribuyendo consistencia en las respuestas y en los correlatos de los puntajes a alguna cualidad, atributo o rasgo de personas u otros objetos a ser evaluados. Este uso apunta al hecho que la interpretación de los puntajes es (o debería ser)

construida para explicar y predecir (o, menos ambiciosamente, para resumir o al menos ser compatible con) propiedades y relaciones de los puntajes.

Dada esta interpretación basada en el razonamiento cuantitativo, la utilización de estos puntajes para la admisión universitaria (implicaciones para la acción) debería estar apoyada en evidencia valorativa y estadística de que (a) tales habilidades de razonamiento tienen relación con o facilitan el aprendizaje en la universidad (relevancia); (b) los puntajes realmente predicen éxito en el primer año universitario (utilidad); (c) los impactos desfavorables contra mujeres u otros grupos minoritarios, no se deben a un contenido sesgado en favor de los hombres u otras mayorías, o a alguna otra variación en la prueba que sea irrelevante para el constructo, sino que refleja diferencias intrínsecas a los grupos que sí son relevantes para el constructo y que afectan el desempeño cuantitativo en la prueba (valoración de las consecuencias de los efectos secundarios o colaterales). Por consiguiente, los puntos centrales en la validación de pruebas son el significado, la relevancia y la utilidad de los puntajes, las implicaciones de valor de los puntajes como base para la acción, y el valor funcional de esos puntajes en términos de las consecuencias sociales que su uso pudiera tener.

## MÚLTIPLES LÍNEAS DE EVIDENCIA PARA LA VALIDEZ UNIFICADA

La validez es un concepto unitario, a pesar de que hay diferentes fuentes y combinaciones de evidencia para apoyar las inferencias hechas a partir de los puntajes. La validez siempre se refiere al grado en que la evidencia y la teoría sustentan la pertinencia y relevancia de las interpretaciones y acciones realizadas a partir de los puntajes de las pruebas. Además, aunque hay muchas maneras para reunir evidencia favorable a una inferencia en particular, estas maneras no son otras que las propias de los métodos de la ciencia. Las inferencias son hipótesis; y la validación de inferencias es la prueba de las hipótesis. Empero, no se trata simplemente de probar hipótesis, sino, en términos generales, de probar teorías, pues la fuente, el significado y el sentido de las hipótesis relativas a puntajes provienen de las teorías

sobre la interpretación del significado de los puntajes en que dichas hipótesis se encuentran enraizadas. En consecuencia, la validación de pruebas está básicamente ligada, tanto a la teoría como a la información cuantitativa. Por lo mismo, la validación de pruebas involucra todos los medios experimentales, estadísticos y filosóficos normalmente vinculados a la validación de hipótesis y teorías científicas. Lo que sigue amplía aún más estos dos puntos básicos, a saber, que la validez es un concepto único pero con diversos matices; y que la validación es una indagación científica sobre el significado de los puntajes.

**Fuentes de evidencia de validez**: Las fuentes fundamentales de evidencia de validez no son, en modo alguno, ilimitadas. La verdad es que si a uno le preguntaran dónde buscar esas fuentes de evidencia sólo encontraría alrededor de una media docena de estrategias principales de investigación y otras formas relacionadas de evidencia. La cantidad de formas es arbitraria, desde luego, porque éstas pueden clasificarse en diferentes formas y categorías establecidas en distintos niveles de generalidad. Pero alrededor de una media docena de categorías como las siguientes, ofrece posibilidades para destacar semejanzas y diferencias entre enfoques de validación:

1. Evaluar la relevancia y representatividad del contenido de la prueba en lo relativo al dominio de comportamiento o desempeño, sobre el cual se harán inferencias o predicciones.
2. Examinar la relación entre respuestas a las tareas, ítemes o partes de la prueba, es decir, delinear la estructura interna de las respuestas.
3. Explorar las relaciones entre las respuestas a las tareas, ítemes o partes de la prueba, es decir, establecer la estructura externa de la prueba.
4. Investigar directamente la forma en que los individuos se enfrentan a los ítemes o a tareas, tratando de aclarar los *procesos* que subyacen a las respuestas en cada ítem y al desempeño en las distintas tareas.
5. Investigar hasta qué punto estos procesos y estructuras se mantienen uniformes o cambian a lo largo del tiempo o según grupos y contextos; esto es: asegurar que el

poder de generalización (y los límites) de la interpretación y el uso de la prueba son apropiados para el constructo y para los contextos en juego.

**6.** Evaluar el grado en que los puntajes muestran variaciones teóricamente apropiadas o esperadas como consecuencia de la enseñanza o de otras intervenciones, o como resultado de la manipulación experimental del contenido y de las condiciones.

**7.** Estimar las implicaciones de valor y las consecuencias sociales de interpretar y utilizar los puntajes de la manera propuesta, escrutando no sólo los resultados esperados sino también los efectos colaterales imprevistos. En particular, evaluar hasta qué punto una consecuencia desfavorable de la prueba se origina en alguna fuente de invalidez para los puntajes, como podría ser una variación irrelevante en la prueba (desde luego, lo mejor es poder descartar que esto haya ocurrido).

El principio rector de la validación de pruebas es que el contenido de la prueba, su estructura externa e interna, sus consecuencias sociales, los procesos de respuesta operativa, el grado de poder de generalización (o su ausencia) y las variaciones de puntaje como función de las intervenciones y manipulaciones tengan sentido en función del atributo o rasgo (más ampliamente, el constructo) que se mide a través de los puntajes. La evidencia de investigación que no corresponda con la teoría pone en cuestión la validez de la medición o la validez del constructo, o la de ambos, siempre y cuando la validez de la investigación misma no esté en duda.

En el pasado, algunas de estas formas de evidencia de validez, o alguna combinación de ellas, ha sido colocada en una posición especial como un cierto *tipo* de validez. Pero como todas estas formas de validez reposan fundamentalmente en el uso e interpretación válidos de los puntajes, no es un tipo de validez lo que debe determinar el enfoque de la validación, sino la relación entre la evidencia y la inferencia que de ella se desprende. Es decir, se debe buscar evidencia para apoyar (o socavar) tanto la interpretación y uso que se haga de la prueba y sus

resultados, como para descartar otras interpretaciones plausibles. En esta búsqueda, las variedades de evidencia no son alternativas que compiten entre sí, sino más bien complementos unas de las otras. Esta es la razón principal por la cual se reconoce hoy día a la validez como un concepto unitario (APA, 1985) y por la que se tiene a los otros tipos históricos de validez como limitantes en alguna forma.

## TIPOS TRADICIONALES DE VALIDEZ Y SUS LIMITACIONES

Por lo menos desde comienzos de los años cincuenta la validez se ha dividido en tres tipos, cada uno de los cuales, a su vez, comprende dos subtipos -o, más específicamente, en tres tipos, uno de los cuales abarca dos subtipos. Éstos son: validez de contenido, validez predictiva, validez concurrente relativa a criterios (o validez de criterio) y validez de constructo. Estos tres tipos tradicionales de validez han sido descritos, con pequeñas variaciones, de la manera siguiente (APA, 1954, 1966):

- La validez de contenido se evalúa mostrando cuán bien el contenido de la prueba muestrea la clase de situaciones o materia sobre la cual habrán de sacarse conclusiones.
- La validez relativa a criterios se evalúa comparando los puntajes con una o más variables externas (los criterios) que pueden suministrar una medida directa de la característica o comportamiento en cuestión.
- La validez de predicción indica hasta dónde el nivel futuro de desempeño de un individuo, en un criterio dado, puede ser predicho a partir del desempeño en una prueba anterior.
- La validez concurrente indica hasta qué grado el puntaje en la prueba estima la situación presente de un individuo en el criterio.
- La validez de constructo se evalúa investigando qué cualidades mide una prueba, esto es, determinando el grado hasta el cual ciertos conceptos explicativos o constructos dan cuenta del desempeño en la prueba.

Estas concepciones sobre la validez, con algunas modificaciones importantes, aparecen en los estándares y lineamientos actuales para realizar pruebas. Se dan aquí en su versión clásica o tradicional para ofrecer un punto de referencia contra el cual evaluar el sentido de cambios posteriores. Entre éstos encontramos, por ejemplo, un cambio en el enfoque de la validez de contenido que va desde el muestreo de situaciones o temas al muestreo de dominios de comportamientos o procesos; y un cambio en la validez de constructo, que va desde estar en contradicción con la validez de contenido y con la de criterio, hasta incluir a estos dos tipos de validez.

Históricamente, no sólo se han hecho distinciones entre los tres tipos de validez, sino que cada tipo ha sido relacionado con distintos objetivos de evaluación (APA, 1954, 1966). Esto era particularmente riesgoso pues implicaba que había propósitos de prueba para los cuales bastaba solamente uno u otro tipo de validez. La validez de contenido, por ejemplo, era considerada conveniente para apoyar supuestos relativos al nivel de desempeño actual de un individuo en un universo de tareas o situaciones; la validez relativa a criterios, para fundamentar la situación presente o futura de alguien en algunas variables significativas distintas de la prueba; y la validez de constructo, en apoyo de suposiciones acerca de la medida en que un individuo posee una cualidad que se refleja en su desempeño en la prueba.

Sin embargo, por razones que se expondrán detalladamente más adelante (ver también Messick, 1989a, 1989b), ni la validez de contenido ni la relativa a criterios, por sí solas, bastan para sustentar un propósito de evaluación; en tanto que la generalidad de la validez de constructo debe ser puesta a tono con la relevancia, la utilidad y las consecuencias de la interpretación y el uso de los puntajes en contextos de aplicación determinados. Si se comparan estos tipos de validez con las formas de evidencia esbozadas anteriormente, uno fácilmente puede discernir sobre qué forma de evidencia se apoya cada tipo de validez, pero también qué es lo que cada una deja fuera. El resto de esta sección se dedica a subrayar las propiedades sobresalientes y las limitaciones críticas de cada uno de los “tipos” tradicionales de validez.

**Validez de contenido**: Según la explicación clásica, la validez de contenido se fundamenta en juicios especializados acerca de la relevancia del contenido de una prueba respecto del contenido de un dominio o campo del comportamiento que sea de interés, así como respecto de la medida en que los ítemes o tareas cubren ese dominio. Por ejemplo, la relevancia y representatividad de los ítemes en una prueba de química pueden estar dadas por su relación con material usualmente cubierto por el currículo y el libro de texto; en una prueba para seleccionar personal de conserjería, la relevancia y la representatividad de los ítemes pueden estar dadas por las propiedades y funciones del puesto que se hayan revelado en un análisis; y en una prueba de personalidad, pueden estar dadas por los comportamientos y situaciones propias de una particular teoría sobre rasgos de personalidad. De este modo, lo esencial en el concepto de la llamada validez de contenido es que los ítemes de una prueba son muestras de un dominio de comportamiento o universo de ítemes sobre los cuales se harán inferencias o predicciones.

Según Cronbach (1980), "Lógicamente,… la validez de contenido se establece sólo en la construcción de pruebas al especificar un dominio de tareas y al hacer un muestreo riguroso. La inferencia que se hace después hacia el dominio puede ser netamente deductiva." (p. 105) Pero tal inferencia no va desde la muestra de ítemes al dominio de conocimientos, habilidades o de cualquier otro constructo relacionado, sino al dominio de *tareas* consideradas pertinentes para ese constructo. En este sentido, es útil distinguir entre el *dominio* de conocimiento u otros constructos y el *universo* de tareas relevantes (Messick , 1989b). Los juicios sobre la relevancia son importantes para especificar el universo de tareas; éstos, así como los juicios sobre la representatividad, contribuyen a apoyar las inferencias hechas desde la prueba sobre el universo de tareas. Con todo, estas inferencias deben ser matizadas, y hay que reconocer que, si bien la prueba es una muestra del universo de tareas, lo que hace es también acomodar dicha muestra en un formato de prueba; con lo cual eleva el número de efectos de contexto o las variaciones metodológicas por causas irrelevantes frente al dominio de desempeño. Sobre estos efectos insistiremos luego. En cualquier caso, las inferencias sobre el grado en que la muestra o el

universo de tareas se conectan o corresponden con el dominio de conocimiento, de habilidades u otro atributo, requieren no de juicios sobre el contenido sino más bien de evidencia de constructo.

La inconsistencia o confusión entre el dominio del constructo y el universo de tareas ha sido históricamente evidente, sobre todo en relación con la forma de evidencia ofrecida para sustentar la relevancia y representatividad. A lo largo de los años, se ha conceptualizado la validez de contenido de tres maneras muy relacionadas, pero diferentes: en función de cuán bien el *contenido* de la prueba muestrea el contenido del dominio de interés (APA, 1954, 1966); como el grado en que los *comportamientos* manifestados en una prueba de desempeño son representativos de los comportamientos propios del dominio de desempeño especificado (APA, 1974); y, finalmente, como el grado en que los *procesos* utilizados por el examinando para obtener sus respuestas es representativo de los procesos que están en la base del dominio de las respuestas (Lennon, 1956). En la práctica, sin embargo, la evidencia relacionada con el contenido toma frecuentemente la forma de juicios profesionales consensuados sobre la relevancia de contenido de los ítemes (que presumiblemente tienen validez de constructo) para el dominio especificado; juicios que, además, también se refieren a la representatividad del contenido de la prueba con respecto al contenido del dominio. Pero el caso es que las inferencias sobre comportamientos necesitan evidencia de que los comportamientos y el desempeño son consistentes, y no sólo juicios de contenido, mientras que las inferencias relativas a los procesos requieren evidencia relacionada con el constructo (Loevinger, 1957).

Para precisar un poco más sobre el tipo de evidencia de validez que se desconoce o se deja de lado, hay que decir que la validez de contenido por sí misma no tiene relación con procesos de respuesta, con la estructura externa e interna de la prueba, con diferencias de desempeño según grupos y ambientes, con la sensibilidad de los puntajes ante la intervención experimental, o con consecuencias sociales. De este modo, se ve que la validez de contenido arroja evidencia de juicio en respaldo de la relevancia del dominio y de la representatividad del contenido del instrumento, antes que evidencia en respaldo de las inferencias que deben hacerse

de los puntajes de la prueba. La consistencia en las respuestas y en los puntajes en la prueba ni siquiera se considera entre las explicaciones usuales de la validez de contenido. Sin duda, algunas especificaciones de prueba sí se refieren a niveles de conocimiento o a procesos de respuesta deseados. Pero la validez en tales casos, al ser inferida no del contenido de la prueba sino de las consistencia en las respuestas de la prueba y sus correlatos, está claramente relacionada con el constructo.

En un nivel fundamental, entonces, la así llamada validez de contenido no califica como validez, aunque las consideraciones sobre la relevancia y la representatividad del contenido sí califican, y deberían influir, por ello, en las inferencias hechas sobre el puntaje a partir de otra evidencia. Es decir, que la relevancia y la representatividad del contenido de la prueba deben ser consistentes con el rango o generalidad de la interpretación que se haya hecho del constructo. Y viceversa: la generalidad de la interpretación que se haya hecho sobre el constructo debería estar limitada por la relevancia y la representatividad del contenido de la prueba, a menos que pueda sustentarse en otra evidencia que permita la generalización, como podrían ser las correlaciones externas o patrones de factores con medidas de constructo más amplias.

Por lo demás, el problema generalizado de la variación irrelevante en la prueba, sobre todo el de la variación en el método, simplemente no es afrontado en el marco de la validez de contenido, aun cuando la variación irrelevante obstaculiza los juicios relativos a la relevancia del contenido. La variación en el método se refiere a todos los efectos sistemáticos asociados con un procedimiento particular de medición y que son ajenos al constructo puntual que se mide (Campbell & Fiske, 1959). Aquí se consideran todos los efectos contextuales o factores situacionales (como la atmósfera evaluativa) que influyen sobre el desempeño en la prueba y que son distintos del desempeño propio a un determinado dominio (Loevinger, 1957). Por ejemplo, los expertos podrían considerar unos ítemes claramente relacionados con conocimiento o razonamiento como muy adecuados para la resolución de problemas de un dominio, pero los ítemes podrían realmente estar midiendo comprensión de lectura (o esto además de lo otro). O

bien, podría ocurrir que unos ítemes de selección múltiple preparados para medir conocimientos o habilidades contengan distractores tan obvios, que lo reflejado sea realmente simple “habilidad para la prueba” o sentido común elemental. En otro caso, los puntajes atribuidos de manera subjetiva a la capacidad de persuasión de un texto escrito podrían reflejar un buen dominio de puntuación y gramática, o estar influenciados por el volumen de la muestra escrita presentada.

No hay duda de que la variación irrelevante en la prueba contribuye, junto con otros factores, a la debilidad última de la validación tradicional de contenido. Esto quiere decir que el juicio de los expertos es falible y que puede aprehender imperfectamente la estructura del dominio, o representar inadecuadamente la estructura de la prueba, o ambos. Por tanto, tal como se indicó anteriormente, la validez de contenido por sí sola es insuficiente para respaldar algún propósito de prueba, con la posible excepción de muestras de prueba que sean realmente muestras de dominio observadas en condiciones naturales. Aun aquí, sin embargo, la legitimidad de la muestra de una prueba que se considere representativa del dominio del constructo debe, en última instancia, sustentarse en evidencia relacionada con el constructo. La forma de salir de este callejón sin salida es evaluar (e informar) el juicio de expertos sobre la base de otra evidencia; evidencia que versará sobre la estructura del dominio de comportamientos que se esté considerando, así como sobre la estructura de las respuestas en la prueba, es decir, sobre evidencia relacionada con el constructo.

**Validez relativa a criterios:** En contraste con la validez de contenido, la validez relativa a criterios se basa en el grado de correlación empírica existente entre los puntajes de la prueba y los puntajes de criterio. Esta correlación sirve luego como una base que permite utilizar los puntajes de la prueba para predecir la posición de un individuo en una medida de criterio de interés, como el promedio ponderado en la universidad o el éxito en un empleo. Por lo mismo, la validez relativa a criterios solamente hace énfasis en partes seleccionadas de la estructura externa de la prueba. El interés de este tipo de validación no está puesto, por lo general, en el patrón de relaciones de los puntajes en las pruebas con otras medidas, sino que se concentra de manera más

restringida en arrojar luz sobre relaciones seleccionadas con medidas propuestas como criterios para un fin aplicado específico en un entorno dado. La consecuencia de esto es que hay tantas posibilidades de validez relativa a criterios como haya medidas y entornos de criterio, y la medida en que una correlación de criterio puede generalizarse a distintos entornos y momentos se ha convertido en una cuestión empírica importante, aunque polémica (Schmidt, Hunter, Pearlman, & Hirsch, con comentario por Sackett, Schmitt, Tenopyr, Keho, & Zedek, 1985).

En lo esencial, pues, la validez relativa a criterios no está relacionada con evidencia alguna, a no ser por las correlaciones específicas entre prueba y criterio o, de manera más general, por el sistema de regresión que vincula el criterio con los puntajes de predicción. Sin embargo, los puntajes de criterio son *medidas* que deben ser evaluadas como cualquier otra medida. Ellos también pueden tener deficiencias para captar el dominio del criterio de interés y pueden estar contaminados por una variación irrelevante -como en las calificaciones de los supervisores, por ejemplo, que a menudo están distorsionadas por la percepción selectiva, el efecto de 'halo' y otros sesgos. En consecuencia, las medidas de criterio potencialmente deficientes o contaminadas, no pueden servir como estándares inequívocos para validar una prueba, tal como se establece en la definición intrínseca de un enfoque 'de criterios' sobre la evaluación.

Así, como se señaló anteriormente, la validez relativa a criterios, *per se,* es insuficiente para sustentar cualquier propósito de prueba, con la posible excepción (si bien, bastante improbable) de las pruebas de predicción que gozan de alta correlación con criterios completos incontaminados. Aun aquí, no obstante, la legitimidad de la medida de criterio como una muestra del dominio del criterio, es decir, el grado hasta el cual logra aprehender el constructo de criterio, en última instancia debe reposar sobre la evidencia relacionada con el constructo y sobre argumentos racionales (Thorndike, 1949). Se da aquí una paradoja: que los criterios, siendo medidas que deben evaluarse como pruebas, no pueden servir como los estándares para evaluarse a sí mismos. La forma de salir de este callejón sin salida es evaluando ambos, la medida del

criterio y la prueba, en relación con las teorías sobre el constructo propias del dominio del criterio.

**Validez de constructo:** En principio, pero también en la práctica, la validez de constructo se fundamenta en una integración de cualquier evidencia que se refiera a la interpretación o significado de los puntajes –esto incluye a la evidencia de contenido y a la relativa a criterios, las que pasan a ser consideradas como aspectos de la validez de constructo. Dentro de la validación de un constructo, el puntaje en la prueba no es equivalente al constructo que intenta aprehender, ni tampoco se considera que defina al constructo, como en un estricto operacionalismo (Cronbach & Meehl, 1955). La medida se ve, más bien, como una más entre un amplio conjunto de indicadores del constructo. Las relaciones empíricas convergentes, que reflejan características comunes entre tales indicadores, pueden interpretarse como la acción del constructo hasta el punto en que la evidencia discriminante descarta la intromisión de constructos alternativos como hipótesis competidoras plausibles.

Existen dos amenazas importantes para la validez de constructo: una es la *subrepresentación del constructo*, a saber, que la prueba es demasiado específica y excluye dimensiones o facetas importantes del constructo; la otra es una variación irrelevante en el constructo, es decir, la prueba es demasiado amplia y contiene exceso de variación confiable, asociada con otros constructos diferentes y con variaciones en el método, lo que hace que los ítemes sean más fáciles o más difíciles para algunos de los que toman la prueba, de una forma irrelevante para la interpretación del constructo. En resumen, puede decirse que la validez del constructo comprende la evidencia y los razonamientos justificativos que dan apoyo a la confiabilidad de la interpretación de los puntajes; y esto, en función de los conceptos explicativos que dan cuenta tanto del desempeño en la prueba como de las relaciones de los puntajes con otras variables.

En su forma más simple, la validez de constructo suministra la evidencia básica para la interpretación de los puntajes. Como integración de evidencia para el significado de los puntajes, se aplica a cualquier interpretación de puntajes, no sólo a las que estén relacionadas con los llamados “constructos teóricos”. Visto así, no hay por qué discutir acaloradamente si se necesita o no la evidencia de constructo porque el puntaje en cuestión podría no referirse a un constructo teórico, como por ejemplo, al argumentar que la aptitud o idoneidad del maestro (el conjunto de cosas específicas que los maestros conocen, hacen o en las que creen) “no parece ser un constructo teórico” (Mehrens, 1987, p. 215). No tiene importancia si uno opina que la idoneidad, el conocimiento, la habilidad o la creencia son constructos. Si se interpretan los puntajes en función de tales aspectos, entonces hay que aportar evidencia convergente y discriminante de que los individuos con altos puntajes demuestran competencia en un determinado dominio (es decir, haciéndolo posible mediante conocimiento y destrezas) al realizar una tarea, que es lo opuesto a contestar los ítemes ya sea por simple memorización, habilidad adquirida para contestar por tanteo o mero sentido común. Más importante todavía es ser cautelosos al interpretar los puntajes bajos como falta de competencia; antes, deben descartarse un número de hipótesis competidoras plausibles para explicar el bajo rendimiento, como son la ansiedad, la fatiga, la baja motivación, el conocimiento limitado de la lengua o alguna condición de minusvalía. De hecho, este proceso de descartar otras hipótesis plausibles es el sello característico de la validación de constructo (Messick, 1989b).

Más que cuestionar el fundamento del constructo para la interpretación de un puntaje particular, lo más prudente es admitir simplemente la omnipresencia de los constructos; y reconocer que lo que a menudo es objeto de disputa es el grado hasta el cual son explícitamente teóricos, o sea, hasta qué punto están fundamentados en una teoría detallada o incluidos en una red nomológica de relaciones esperadas. Sin embargo, en la medida en que una interpretación de puntajes tenga sólo un sustento teórico pequeño o muy vago, la validación de constructo se hará aún más necesaria, pues ella permitirá aclarar y sustentar el significado de los puntajes.

Casi cualquier tipo de información relativa a una prueba puede contribuir a comprender el significado de los puntajes, pero tal contribución resulta más valiosa si se evalúa explícitamente el grado en que la información se adecua con el fundamento teórico sobre el que se asienta la interpretación del puntaje. Históricamente, el énfasis primordial en la validación de constructo se ha puesto sobre las estructuras externas e internas de la prueba; en otras palabras, en la valoración de patrones teóricamente esperados de relaciones entre puntajes de ítemes o entre puntajes de prueba y otras medidas. Para esclarecer el significado de los puntajes, tal vez son más útiles los estudios sobre diferencias de desempeño esperadas en el tiempo (como los puntajes que muestran un incremento en la capacidad de controlar impulsos, desde la niñez hasta la adultez temprana); los estudios que muestran diferencias entre distintos grupos y entornos (como cuando se comparan las estrategias empleadas por individuos inexpertos frente a las de los expertos, para evaluar la capacidad de resolver problemas en un dominio determinado, o cuando, para evaluar la creatividad, se comparan las producciones de individuos auto-motivados con las de individuos dirigidos por otros); y los que muestran diferencias que surgen en respuesta a tratamientos y manipulaciones experimentales (por ejemplo, un incremento en los puntajes de conocimiento que sea resultado de haber recibido enseñanza específica a un dominio, o una mejora en la motivación para el desempeño que sea resultado de una mayor cantidad de riesgos y beneficios). Posiblemente lo que mejor esclarece el significado de los puntajes es explorar y modelar los procesos subyacentes a las respuestas de la prueba (verbigracia, a través de procesos de “pensar en voz alta” mientras se realiza la tarea), que, con los avances de la psicología cognitiva, son técnicas cada vez más accesibles e influyentes (Snow & Lohman, 1989).

Según se indicó antes, la validez de constructo, aparte de depender de estas formas de evidencia, también está ligada a la relevancia y representatividad del contenido, así como a relaciones con criterios. Así pues, tal información sobre el rango y los límites del contenido cubierto, y sobre comportamientos relativos a criterios que los puntajes predicen, contribuye claramente a la interpretación de los puntajes. En el segundo caso, las correlaciones entre puntajes y medidas de criterio (vistas en el contexto más amplio de otra evidencia que apoye el

significado de los puntajes) contribuyen a la validez de constructo, tanto del predictor como del criterio. En otras palabras: las relaciones empíricas entre los puntajes del predictor y las medidas de criterio deben tener sentido, en términos teóricos, en función de lo que la prueba predictora está supuesta a medir y lo que se supone que el criterio representa (Gulliksen, 1950).

De cualquier manera, estas tres formas tradicionales de validez, tomadas juntas, hacen referencia explícita a todas las formas de evidencia de validez mencionadas anteriormente, excepto una. Esto es así a pesar de que los referentes tanto de la validez de contenido como de la validez referida a criterios son singulares y *ad hoc*, y gracias a que los referentes de la validez de constructo son bastante amplios. Estas formulaciones tradicionales únicamente excluyen la evidencia de validez que trata sobre las consecuencias de la interpretación y uso de las pruebas.

Resulta irónico que, por tanto tiempo, se haya puesto tan poca atención a las importantes consecuencias de la validación de pruebas; esto es porque en el pasado la validez ha sido concebida, de manera convincente, a partir del valor funcional de las pruebas, es decir, tomando en cuenta cuán bien la prueba hace lo que se quiere que haga (Cureton, 1951; Rulon, 1946). Pero para determinar esto último, uno no sólo debe preguntarse si las consecuencias reales y potenciales de la interpretación y uso de la prueba apoyan los propósitos de la evaluación misma, sino también si son consistentes con otros valores sociales. No obstante, este tipo de evidencia no debe ser considerada de manera aislada, como un cuarto tipo de validez, algo así, digamos, como una “validez de consecuencia”. Más bien, debido a que las consecuencias de los resultados, esperados e inesperados, de la interpretación y uso de las pruebas provienen del significado de los puntajes (y a la vez contribuyen a él), tomar en cuenta las consecuencias sociales de una prueba se convierte en un aspecto más de la validez de constructo (Messick, 1964, 1975, 1980).

De lo antedicho se desprende que la validación de constructo surge de estas muchas fuentes de evidencia a manera de un mosaico de hallazgos convergentes y discriminantes que dan apoyo al significado de los puntajes. Sin embargo, para aplicaciones dadas de la prueba, tal

mosaico de evidencia *general* puede o no incluir información específica pertinente sobre la relevancia de la prueba para el propósito particular para el que se esté aplicando, o sobre la utilidad de ésta para el entorno en que se aplica. Por ello, en algunas circunstancias la evidencia general que apoya la validez de constructo puede requerir el apoyo de evidencia específica sobre relevancia y utilidad. La relevancia del uso de la prueba implica que hay evidencia (incluyendo la de contenido y la relativa a criterios) de que la prueba refleja válidamente procesos o constructos considerados importantes en el dominio de aplicación. Así por ejemplo, la evidencia de relevancia podría implicar que a nivel teórico se vincule la evidencia de contenido con dimensiones del desempeño en un dominio, derivadas del análisis de un trabajo o tarea, o que esta vinculación se dé a nivel empírico con los puntajes de ítem o de prueba. La utilidad del uso de una prueba goza de los beneficios relativos a los costos de la prueba misma, frecuentemente en función del grado de correlación entre los criterios de la prueba (Cronbach & Gleser, 1965). En cierto sentido, pues, la amplia generalidad de la validación del constructo puede producir algunas limitaciones, reales o potenciales, en el uso de la prueba, a menos que el mosaico de hallazgos generales del constructo incluya evidencia de relevancia y utilidad en el entorno de aplicación, o hasta que tal evidencia pueda desarrollarse. En otras palabras: al analizar el significado de los puntajes, la evidencia general para la validez de constructo puede no ser suficientemente precisa o específica para garantizar el uso de la prueba para un propósito y entorno particular.

**Balance de las limitaciones propias de los tipos de validez:** Aunque ahora veamos los tres tipos tradicionales de validez como aspectos o facetas de un concepto unitario, existe aún la necesidad de subrayar las fortalezas y, especialmente, las debilidades de esos tipos de validez, para ilustrar mejor la necesidad y la conveniencia de tener una única concepción. La piedra de toque de la validación de contenidos es el juicio de los expertos, quienes especifican la relevancia y representatividad del contenido de la prueba frente a un dominio de contenido. En la validez relativa a criterios, la piedra de toque es la medida del criterio, que sirve como estándar para evaluar la relevancia y la utilidad de los puntajes en la prueba. El problema de fondo, sin

embargo, es que tales piedras de toque no son sólo falibles o están sujetas a error, sino que inducen confusiones, pues el juicio de los expertos no sólo puede ser poco confiable sino que también puede estar prejuiciado, en tanto que los criterios pueden no sólo estar contaminados sino que también pueden ser incompletos. En comparación, cuando se trata de la validación de constructo, la piedra de toque es la evidencia convergente y discriminante, que ratifica el significado de los puntajes y descarta otras posibles interpretaciones. Aunque cualquier evidencia puede ser falible y en algunos casos hasta ser engañosa, el proceso continuo de la validación de constructo intenta evaluar, y tomar en cuenta, la naturaleza y el grado de tal distorsión en el juicio de validez, conforme este evoluciona.

Desde luego, tanto la evidencia convergente como la discriminante suministran una base racional para evaluar las otras dos piedras de toque de carácter dudoso propias de la validez de contenido y de la validez de criterio, tanto al momento de la interpretación como en instancias específicas en que se usa la prueba. En otras palabras: la evidencia relacionada con el constructo es decisiva al delinear los dominios de cada contenido, así como en la conceptualización y evaluación de los criterios aplicados, es decir, justamente en los aspectos de cobertura de dominio y de predicción de criterio que están en el corazón de la validez de contenido y de la validez de criterios tal como han sido tradicionalmente concebidas. Desde este punto de vista, la validez de constructo al momento de interpretar puntajes sirve de base a *todas* las inferencias que se hagan a partir de dichos puntajes, no solamente a las que se relacionan con la significatividad de la interpretación, sino también a las inferencias relacionadas con contenidos y criterios que sirven para tomar decisiones y acciones a partir de los puntajes.

Por otra parte, según se indicó anteriormente, puede ocurrir que el mosaico de evidencia general sobre la validez de constructo, que respalda la interpretación de los puntajes, siga siendo limitado para los usos particulares propuestos para la prueba. En casos así, el mosaico se debe extender para incluir evidencia sobre la relevancia de la prueba para el propósito con que se le aplica, así como sobre su utilidad en el entorno de aplicación. Como se ve, en algunos propósitos

de aplicación, las consideraciones sobre contenidos específicos y criterios seleccionados vuelven a surgir como parte de la validez general de constructo al momento de interpretar el significado de los puntajes. Desde luego, ni la validez de contenido, ni la validez referida a criterios, por sí solas, son suficientes para respaldar la aplicación específica de la prueba: porque, en última instancia, para justificar el empleo de la prueba, se necesita saber el significado de los puntajes tanto con respecto a la prueba misma, como con respecto a criterios (Loevinger, 1957; Thorndike, 1949). Pero en contextos en que la evidencia sobre la validez de constructo es menos específica, esos otros tipos de validez pueden respaldar (o no) las implicaciones que el significado atribuido a los puntajes tiene para la acción, lo que en última instancia justifica un determinado uso de las pruebas.

Si bien en la práctica cada uno de los enfoques tradicionales sobre validación resulta, real o potencialmente, problemático, esto puede ser superado si se les trata de manera conjunta o, mejor, si al momento de hacer una validación de constructo se analiza y asegura que ésta incluya consideraciones sobre contenido, sobre criterios y sobre las consecuencias de la interpretación y uso de las pruebas. Así entonces, la validación de pruebas no puede *depender* de una sola de las seis o siete formas de evidencia discutidas anteriormente, y, siempre y cuando haya evidencia convergente y divergente defendible para respaldar el significado de los puntajes, tampoco *requiere* de una única forma de evidencia. Sin embargo, hay casos en que no se puede desarrollar forma de evidencia alguna: como cuando investigaciones relacionadas con criterios deben desecharse por el pequeño tamaño de las muestras, porque los criterios no son confiables o están contaminados, o porque los rangos de puntajes son muy restringidos. En tales situaciones, debe hacerse énfasis en otra evidencia, especialmente en la relativa a la validez de constructo de las pruebas predictoras y en la relevancia del constructo respecto del dominio del criterio (Guion, 1976; Messick, 1989b). De esta forma, la validez se convierte en un concepto unificado y, la fuerza unificadora es el significado o la interpretación confiable de los puntajes y de sus implicaciones para la acción, esto es, la validez de constructo.

## FACETAS DE LA VALIDEZ UNIFICADA

Lo esencial en la validez unificada es que la conveniencia, el significado y la utilidad de las inferencias hechas a partir de los puntajes son inseparables; y que la fuerza integradora proviene de una interpretación de los puntajes empíricamente fundamentada. No obstante, hablar de la validez como un concepto unificado no implica que ella no pueda ser diferenciada en facetas o aspectos para enfatizar temas y matices que, de otra forma, podrían ser desestimados o pasados por alto, como es el caso de las consecuencias sociales o el papel del significado de los puntajes en el uso aplicado de las pruebas. Estas distinciones sirven para comprender aspectos funcionales de la validez que puedan ser útiles para desenmarañar algunas de las complejidades inherentes a las consideraciones sobre la conveniencia, el significado y la utilidad de las inferencias hechas sobre los resultados. Lo que se necesita, en especial, es dar forma a una evidencia de validez que evite una dependencia indebida en ciertas formas de evidencia, que resalte el papel importante, si bien secundario, de la evidencia de contenido y de la relativa a criterios, como respaldo de la validez de constructo en el uso de pruebas, y que considere también, formalmente, los valores implicados y las consecuencias sociales dentro de la estructura de la validez.

**Significado y valores como maneras para conformar la evidencia de validez.** Una estructura unificada de la validez que cumpla con estos requisitos distingue dos facetas de la validez como concepto unitario (Messick, 1989a, 1989b). Una sirve para justificar la prueba, y tiene que ver ya sea con evidencias que justifican el significado de los puntajes o con las consecuencias que contribuyen a una determinada valoración de éstos. La otra faceta tiene que ver con la función o resultado de la prueba, que puede ser una determinada interpretación o un uso aplicado. Si la faceta que sirve para justificar (es decir, aquélla que consta de un conjunto de evidencias para justificar las implicaciones de significado, o de una justificación de las consecuencias para apoyar las implicaciones del valor atribuido a los puntajes), se cruza con la faceta que trata sobre la función o resultado (es decir, sobre la interpretación o uso de la prueba), se obtiene una tabla de cuatro entradas que ilustra el significado y los valores propios de la interpretación y uso de pruebas, tal como se aprecia en la Tabla 1.

**Tabla 1**

**Facetas de la Validez como una Matriz Progresiva**

| | INTERPRETACIÓN DE LA PRUEBA | USO DE LA PRUEBA |
|---|---|---|
| EVIDENCIAS FUNDAMENTALES | Validez de Constructo (VC) | VC + Relevancia / Utilidad (R/U) |
| CONSECUENCIAS FUNDAMENTALES | VC + Implicaciones de Valor (IV) | VC + R/U + IV + Consecuencias Sociales |

Las cuatro facetas de la validez aquí representadas constituyen aspectos distintos pero interrelacionados de la prueba que juntos buscan responder a las siguientes preguntas: ¿Qué evidencia final justifica la interpretación o significado de los puntajes? ¿Qué evidencia justifica no sólo el significado de los puntajes, sino también la relevancia de éstos para el propósito particular para el que la prueba está siendo aplicada, y la utilidad de los puntajes para el contexto en que se aplica la prueba? ¿Qué brinda credibilidad a las implicaciones valorativas que se derivan de la interpretación de los puntajes, y a las consecuencias que esas interpretaciones tienen para la acción? ¿Qué implica el valor funcional de la prueba en términos de sus consecuencias esperadas e inesperadas?

Consideremos brevemente cada una de las celdas en esta tabla de validez unificada, comenzando con la evidencia fundamental para la interpretación de una prueba. Dado que la validez de constructo ha sido definida como la evidencia y los raciocinios que apoyan la validez del significado de los puntajes, resulta claro que es la validez de constructo la que constituye la evidencia fundamental para la interpretación de una prueba. La evidencia fundamental para el uso de una prueba también está dada por la validez de constructo, pero con la importante

condición de que la evidencia general que apoya el significado de los puntajes, debe incluir desde el comienzo, o debe ser reforzada por, evidencia específica que demuestre la relevancia de los puntajes para el propósito con que se aplicó la prueba, y la utilidad de los puntajes en el contexto de aplicación.

Las consecuencias fundamentales de la interpretación de una prueba están dadas por un reconocimiento de las implicaciones que, en términos de valoraciones, tiene el significado de los puntajes, incluyendo las valoraciones tácitas que están implicadas en el nombre mismo del constructo, aquéllas que subyacen a la teoría a partir de la cual se definen las propiedades y relaciones del constructo y que sustentan el significado de éste, y aquéllas que, de manera todavía más amplia, subyacen a las ideologías (como, por ejemplo, sobre las funciones de la ciencia o sobre la naturaleza de los seres humanos como aprendices) cuyas perspectivas y propósitos informan a las teorías (Messick, 1989b). Así, por ejemplo, un constructo sobre un comportamiento que es variable en contraste con otro repetitivo, podría recibir el nombre de 'flexibilidad *versus* rigidez', y las valoraciones a que este rótulo daría lugar al momento de interpretar los puntajes sería muy diferente de las valoraciones a que hubiera dado lugar si se le hubiera llamado "confusión *versus* control". Del mismo modo, un constructo, y las mediciones con él asociadas, que fuera denominado "inhibido *versus* impulsivo", tendría implicaciones de valor muy distintas que si hubiera sido interpretado como "autocontrolado *versus* expresivo".

Muchos constructos como competencia, creatividad, inteligencia o extroversión tienen múltiples y discutibles implicaciones de valor, que pueden encontrar apoyo, o no, en función de las propiedades de las mediciones con ellos asociadas. Al interpretar una prueba, un tema importante es si las implicaciones teóricas o características y las implicaciones de valor son conmensurables, pues las implicaciones de valor no están subordinadas al significado de los puntajes, sino que son parte integral de éste. Así, para poder dejar en claro que la interpretación de los puntajes es necesaria para poder apreciar las implicaciones de valor y viceversa, la celda

sobre las consecuencias fundamentales para la interpretación de una prueba debe incluir tanto la validez de constructo como las implicaciones de valor del significado de los puntajes.

Por último, las consecuencias fundamentales del uso de una prueba están dadas por consideraciones sobre las consecuencias tanto reales como potenciales de la aplicación de una prueba determinada. Una aproximación para evaluar posibles efectos colaterales es comparar los riesgos y beneficios del uso propuesto para la prueba, con los pros y los contras de otras alternativas o contrapropuestas. De esta manera, al emplear múltiples perspectivas sobre el uso propuesto para la prueba, los valores (distintos y a veces hasta contrapuestos) implicados en cada propuesta quedan expuestos para ser examinados y debatidos abiertamente (Churchman, 1971; Messick, 1989b). Cuando de lo que se trata es de evaluar estándares de desempeño, las contrapropuestas para un determinado uso de una prueba pueden involucrar el uso de distintas técnicas de medición, como observaciones o portafolios. O, cuando se quiere evaluar los niveles de productividad, las contrapropuestas pueden estar orientadas a satisfacer un mismo propósito en formas distintas, como por ejemplo, mediante capacitación y no mediante selección.

Lo importante no es únicamente si las consecuencias sociales de la interpretación y el uso de una prueba son positivas o negativas, sino cómo se originaron esas consecuencias y qué las determinó. En especial, no es que las consecuencias sociales adversas de una prueba hagan su uso inválido, sino que tales consecuencias no deben originarse en alguna fuente de invalidez de la prueba, como es caso de una variación irrelevante en el constructo. Y, de nuevo, puesto que se reconoce que la ponderación de las consecuencias sociales de una prueba a la vez supone y contribuye a generar evidencias para sustentar el significado de los puntajes, la relevancia, la utilidad y los valores asociados a ella, esta celda debe incluir consideraciones sobre la validez, la relevancia y la utilidad del constructo, así como sobre las consecuencias sociales y valorativas de éste.

Así, la validez de constructo aparece en todas las celdas, algo muy adecuado, porque la validez de constructo del significado de los puntajes es la fuerza integradora que unifica todo lo relativo a validez en un concepto unitario mayor. Distinguir las facetas que reflejan la justificación y la función de una prueba hace evidente, al mismo tiempo, la necesidad de enfatizar los distintos aspectos de la validez de constructo, además del mosaico general de evidencia. Esto implica que, según nos movamos de una celda a otra en la tabla, el énfasis pasará de consideraciones sobre la evidencia para la interpretación del constructo *per se*, a consideraciones sobre la evidencia que sirve de sustento para un uso racional de la prueba; y de ahí a consideraciones sobre las consecuencias valorativas de la interpretación de los puntajes, que sustentan acciones posteriores; y, finalmente, a consideraciones sobre las consecuencias sociales, o el valor funcional, del uso de una prueba. Conforme se van agregando énfasis distintos a cada una de las facetas de la validez de constructo, vemos que lo que en un primer momento aparecía como una simple clasificación se parece más a una matriz progresiva, como la que se aprecia en las celdas de la Tabla 1.Una de las implicaciones de formular esta matriz progresiva es que tanto el significado como la valoración, así como el uso y la interpretación de una prueba, se entremezclan en el proceso de validación. Así, la validez y la valoración se convierten en un solo imperativo, no dos, y la validación de una prueba conlleva tanto la ciencia como la ética de la medición. 

# Referencias[3]

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin*, 51 (2, Part 2).

American Psychological Association. (1966). *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1974). *Standards for educational and psychological tests*. Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.

Churchman, C. W. (1971). *The design of inquiring systems: Basic concepts of systems and organization*. New York: Basic Books.

Cronbach¸ L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd. Ed., pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L. J. (1980). Validity on parole: How can we go straight? *New directions for testing and measurement -- Measuring achievement over a decade* -- Proceedings of the 1979 ETS Invitational Conference (pp. 99- 108). San Francisco: Jossey-Bass.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana, IL: University of Illinois Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.

[3] Sobre este tema, hay una extensa bibliografía en: S. Messick (1989). Validity. En R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp- 13-103), New York: Macmillan.

710 Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (1st ed., pp. 621-694). Washington, DC: American Council on Education.

Guion, R. M. (1976). Recruiting, selection, and job placement. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 777-828). Chicago: Rand McNally.

Gulliksen, H. (1950). Intrinsic validity. *American Psychologist*, 5, 511-517.

Lennon, R. T. (1956). Assumptions underlying the use of content validity. *Educational and Psychological Measurement*, 16, 294-304.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, 3, 635-694 (Monograph Supplement 9).

720 Mehrens, W. A. (1987). Validity issues in teacher licensure tests. *Journal of Personnel Evaluation in Education*, 1, 195-229.

Messick, S. (1964). Personality measurement and college performance. *Proceedings of the 1963 Invitational Conference on Testing Problems* (pp. 110-129). Princeton, NJ: Educational Testing Service. [Reprinted in A. Anastasi (Ed.). (1966). *Testing problems in perspective* (pp. 557-572). Washington, DC: American Council in Education.]

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 30, 955-966.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, 35, 1012-1027.

730 Messick, S. (1989a). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, 18 (2), 5-11.

Messick, S. (1989b). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Rulon, P. J. (1946). On the validity of educational tests. *Harvard Educational Review*, 16, 290-296.

Schmidt, F. L., Hunter, J. E., Pearlman, K., & Hirsh, H. R., with commentary by Sackett, P. R., Schmitt, N., Tenopyr, M. L., Keho, J., & Zedeck, S. (1985). Forty questions about validity generalization and meta-analysis. *Personnel Psychology*, 38, 697-798.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for
740 educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263-331). New York: Macmillan.

Thorndike, R. L. (1949). *Personnel selection*. New York: Wiley.